Personalized Web Directory :A Knowledge Discovery Approach

Madhavi S. Darokar (Researcher) , Prof. Mansi Bhonsle , G.H. R. C.E. M. Pune, India.

Abstract

The World Wide Web is a rich source of information. The number of users accessing web sites is increasing every day. As the growth of web is tremendous due to which it is not possible to retrieve the online information easily because of ”information overload” problem. To address this problem, personalization is used, which focuses on the retrieval to meet the user-specific information. For effective and efficient handling the web mining techniques provides personalized contents at the disposal of users.

Web usages mining is an area of Data mining dealing with the extraction of interesting knowledge from the World Wide Web. So We are building here ”Knowledge Discovery Framework” for the construction of community specific web directories by applying personalization to web directories. The hierarchical structure of the web pages on the web categories into specific theme of user interest is called as web directories which is generally constructed manually by human experts. Instead of that the Cluster analysis, which deals with the organization of a collection of objects into cohesive groups of interest can play a very important role for automation of this process.

Key terms

Clustering, Personalization , Pattern Analysis, Machine Learning,Web Directory , Web Logs, Web Usages Mining.

1. Introduction

The tremendous amount of data is available on the World Wide Web. Now a days it is cumbersome task to find the relevant data from it because of information overload problem as it’s size is increasing continuously. The main objective of this paper is to construct web directory to reduce this problem with the help of personalization. The content of the web is organized into thematic hierarchy called as web directories which corresponds to listing of those topic in which user is interested. There are some real web directory like DMOZ(Open Directory Project) and Yahoo directory. Here user can get the interested information by searching inside the directory from broad category and gradually narrowing down until they get the thematic contents .So the user has to check deep inside the directory until they get satisfied information on the web. To alleviate this problem we can apply here personalization to web directories using web usages data. A aggregate user community model is constructed from browsing behavior of web in the form of web usages data, for personalization of services on the web. This can be achieved by applying pattern analysis on the web usages data which are in the form of web log data at the server side. The clustering and probabilistic approaches are used for the pattern analysis to build community specific personalized web directory. As the web data has high degree of thematic diversity (increased dimensionality and semantic incoherence), we are creating a knowledge discovery framework for construction of community web directory by applying personalization to web directory which will become automatic machine learning process[3].

1.1 Existing Systems

There are some systems which are full automation of personalized process and employs machine learning methods. One of such system is ”Montage System”which creates personalized portal by applying number of heuristic metrics to web usages data such as the interest in a page or a topic, the probability of revisiting a page, etc. It consists of link to the number of pages the user has visited which are also organized like ODP[4]. The another system is ”Power Bookmark System” which collects the bookmark information of the single user such as frequently visited pages and query results returned by search engine. This system usages text classification technique for categorization of web pages to specific folder. The problem of the system is adoptability is only to single user views and not construct aggregate model for the user and also scalability of the classification methods they use[2]. A Web directory, such as Microsoft (www.microsoft.com) and the Open Directory Project (ODP) (dmoz.org) and Yahoo! the personalized search is possible but directory is not personalized[3]. In this directory web pages are explicitly assigned manually to categories of directory.

1.2 Deficiency in the existing systems

1. Web is a goldmine of information but the ”information overload” becomes frustrating phenomenon to the web users so requires personalized services for information retrieval on the web.

2. Due to tremendous growth in size of web at the current state web has not achieved the goal to navigate

information to the particular user.

3. Web pages are categorized manually, hence limited topic coverage.

4. Because of the size its complexity gets increased results in difficulty of appropriate navigation.

2. Related Work

Many researches have been carried out for the web personalization using web usages data. The pattern discovery from web log data is done by using majority of clustering methods[4]. This method is used to divide the data into groups which are different from each other. Clustering is basically used for groping people having common interest while browsing the web or the web pages having same content. Actually cluster is categorized into three categories[2].

1. Partitioning Methods

2. Hierarchical Methods

3. Model Based Methods

In partitioning method the algorithms used are Leader, PageGather and Expectation Maximization. A Leader algorithm is used for clustering the user sessions which is represented by the n vector where n is the number of web pages accessed in that session. The value of each vector is represented by weight w where w is count of interest in particular web page of user. This algorithm produces good quality clusters. Also the pattern discovery is made by vector and weight characteristics. But this algorithm has drawback like different set of clusters can be generated depends upon training vector sequence provided as parameter to the algorithm. Another algorithm is PageGather used for clustering which takes user session as input and used for improving web site representation by gathering set of pages which are visited by user from the web log data. It has advantage of producing overlapping cluster of same behavioral browsing of users. But it has drawback of its computation cost because of graph based method where nodes of the graph are web pages and edges between the nodes are the cooccurrence of the web pages[16] .The Expectation Maximization algorithm is also partition based which takes input as user sessions in the form of URL's from web log files and represented using categories of web pages of some topic. This algorithm cluster the user session of a particular group called community. The advantage is it is memory efficient but drawback is computationally expensive. This also include fuzzy clustering algorithm.

The hierarchical method the BIRCH algorithm is used for clustering of user session. The web log data is converted to user session which contains the IP and timestamp .The session are organized here like page hierarchy of the web. It is very efficient and applicable to large volume of the data. But the drawback is it depends upon sequence of user inputs.

The model based method Autoclass, SelfOrganizing Map, Incremental algorithm are used to construct user community model having similar interest in web usages pattern[5]. The user communities are clustered as per their interest in browsing the web. The advantage of Autoclass is that it mathematically sound but computationally expensive. A self Organizing Map has good mapping of high dimensional data but it requires prior specification of the number of clusters. An Incremental algorithm is applicable to large dataset but suffer from scalability.

Thus we can achieve Personalization of web based systems from web usages data. For the experimental setup the log file is collected from proxy server of ISP. The software requirement for our project is JDK 7 and IDE required is Net Beans 7.1 with minimum hardware requirement of 1GB Hard Disk ,512 Ram and Window 7 Operating System with Pentium 3 processor.

3. System Architecture

A aggregate user model is constructed here by collecting data from web proxies as user browse the web and applying some machine learning techniques[15]. The main purpose of the project to construct a community web directory aytomatically which result in operational personalized knowledge. The process of getting from the data to the community Web directories is summarized below.

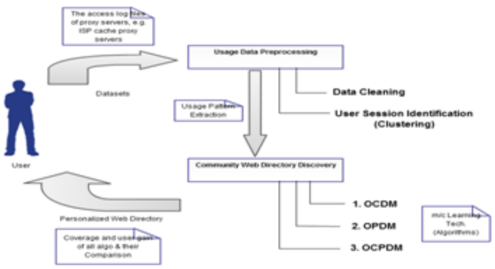

The access log files of proxy servers, e.g. ISP cache proxy servers
Usage Data Preprocessing
Datasets
Data Cleaning
User Session Identification (Clustering)
Usage Pattern Extraction
User
Community Web Directory Discovery
1. OCDM
2. OPDM
3. OCPDM
m/c Learning Tech. (Algorithms)
Personalized Web Directory
Coverage and user gain of all algo & their Comparison

Fig 1: System architecture for web directory personalization.

System architecture basically consists of three blocks
a) Preparation of the Usages data
b) Discovery of community web directory
c)Evaluation of Community Web Directory
These are described as follows.

3.1. Preparation of Usages data

This is the first phase for collection and cleaning of the usages data and user session is identified. The usages data is collected from the log files of Proxy servers. The web log data consists of a date and time stamp, IP address, response type, status code, content type header[17]. But there is no record of user identification to avoid privacy violations. The usages data collected from the log files are semantically diverse and large. The aim of data preparation is to assemble these data into integrated and consistent order so that it can be used for pattern discovery.

The first step is data cleaning. The task is to remove noisy data such as images, multimedia, advertisement banner. Also records with HTTP status error code which means bad request and unauthorized access also should be removed[16]. This make the filtering of log files to keep only the web pages that are directly related to the user thematic behavior.

The second step is the categorization of the web pages into thematic categories which helps to minimize the dimensionality and the diversity of web accessing behavior. It is done mostly by text classification method [9] like yahoo! But it has very low coverage than the web directory constructed using web usages data which is proposed here. We have adopted here a different methodology for web page categorization which depends upon Uniform Resource Locator of the web pages to extract actual domain of user interest from URL then search for the directory which maps within an existing Web Directory, such as ODP[2] instead of finding the content of each web page.

The third step is extraction of user specific session which is conducted for specific time interval. This is sequence of web pages accesses from the same IP address. The access session can be extracted by following ways which forms the main input to the pattern discovery phase.
1) The IP address and date is used to groups the log records.
2) The specific time interval is set for the two records which belong to same IP address in the same session.
3) The categories are grouped together to form a session from same IP address.

3.2. Discovery of community web directory

There is two approaches of unsupervised learning to discover patterns of interest of thematic user session[19]. These are community directory miner and probabilistic latent semantics analysis. A graph of the web directory is constructed for community web directory and community model is denoted byΦ which is used to extract the subset of categories of initial Web Directory.[4] The Pattern Discovery algorithm measure the informativeness of web directory means number of times user visit the category while browsing according to the preferences of the community. This helps to reduce initial web directory and personalized to the interest of the community.

Web directory defination[4]: Consider G=(V,E) be a Web Directory.Let θ be a set of community web categories.We define the Community Web directory as the subgraph G'=(V' , E') of G with the following properties.

- $V' \subseteq \theta \subseteq$ V
- An edge e= (va,vb) ϵ E' from node Va to node Vb is created inG', if {Va , Vb }$\subseteq \theta$ and one of the following conditions is met:

- $\exists$ (Va , Vb) ϵ E

- $\exists$(Va , Va+1,......... Vb+1 , Vb) a path in E, $\wedge$ Va , Va+1,....., Vb+1 $\cap \theta = \Phi$

3.3. Community Web Directory Discovery Algorithms

The user session is given as input to the pattern discovery algorithm and mapped to thematic user session and this mapping is many to one .We are mapping more than one page to same leaf category in the same user session.

Thematic session set[4]: Let u(t1,tf) is a thematic user session,then it's session set $\bar{u}$ (t1,tf)=(li,liϵV) is the set of unique categories in n(t1,tf).
The categories found are categories expressed as the categories of the thematic session set then session and categories are relate to form ancestor categories which can be defined as follows[1].
Thematic session tree: consider the set of thematic session sets U=($\bar{u}$1,$\bar{u}$2.....$\bar{u}$3)and Web directory set of categories V,S Thematic session binary tree is given by relation R=(U,V). If there is pair of($\bar{u}$,vi which means session access a certain category vi ϵ V.

R = { 1 if category exist

0 otherwise

3.4. Objective Community Directory Miner (OCDM)

This is the first machine learning method used for community directory discovery. The simple cluster mining algorithm is used for discovering common behavior by using graph where the vertices correspond to categories[4]. This clustering algorithm discover pattern of common users characteristic features which is known as thematic categories. The edges corresponds to category co-occurrence in thematic categories session sets. In graph the vertices and edges assigned weights which depend upon occurrence and co-occurrence frequencies. Here a user may assign too many communities. The weight of the vertices is denoted by Wi and corresponding categories Vi and the weight of the edge Wij is follows

$Wi = \sum_{k=1}^{n} \frac{R(uk,vi)}{n}$
$Wij = \sum_{k=1}^{n} \frac{R(uk,vi).R(uk,vj)}{n}$
Where n is the total number of the thematic session sets.

The OCDM creates hierarchy of topic categories.Then while accessing web directory each category is mapped onto a set of categories. This is done by updating the weights of the vertices and the nodes in the graph. The result is the construction of a topic tree.If the connectivity of the resulting graph is usually high and threshold is use to reduce the edages of the graph.After this the weighted graph is turned into unweighted and community model θr is constructed from maximal cliques of the unweighted graph.The informativeness of the leaf categories of the initial Web directory are examined then which are not in the clique. Using the OCIA criterion, we compare each such leaf against its closest ancestor that that is included in the θr.

3.5. Objective Probabilistic Directory Miner (OPDM)

This is the second machine learning method which related with the hidden association related within the web usages data. As the OCDM algorithm is based on observable behavior of the user based on web usages data on users interest[4]. But generally users interest and motives are less explicit. The users interests are considered here is motivated here by number of latent factors.So when user might visit pages from a particular category of the web directory not because of same interest but with same motivation and by common subset of them.

As an example a user when navigates through the web pages that belongs to the category"Top/Computer/Companies" means user might motivated by interest in e-commerce or job offers[2][14] . These latent factors are responsible for the association between users. Thus a latent factor is also used for identification of pattern in web usages data and can be used for grouping the data. Thus it provides multidimensional characteristic way of user interest.
Here the each session-category pair is observed due to a latent generative factor that is de noted by the variable zk and hence the generic association between the elements of the pairs is provided.

$$P(Ui, Vj) = \sum p(Zkp(Ui|Zk)p(Vj|Zk)$$

3.6. Objective Clustering and Probabilistic Directory Miner (OCPDM)

This is proposed algorithm for the discovery of community models. This algorithm is a combination of clustering with probabilistic latent semantic analysis[6][21].

The K-means clustering algorithm is applied here for initial communities to avoid the disadvantage of PCDM and non overlapping cluster should be created so that each category belongs to only one cluster. To discover the hidden knowledge each cluster derived from K-means are mapped onto space of latent factors. Thus the community web directory is constructed using latent factor with observable association of web usages data which is better model of users interest.

3.7. *Data independence and Data Flow architecture*

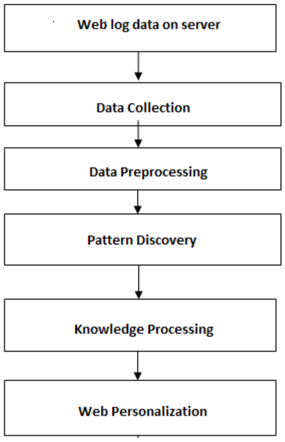


Fig 2: Data Flow for web directory personalization.

4. Results and Discussion

Evaluation of Community Web Directory: The current research approximated the gain of the end user but not taken into account the cost of ”losses” so in that case the users do not find what they are looking for in the personalized directory. This issue requires the evaluation of community Web directories in user studies which we implement in this project . The web log files are collected from the ISP and the proposed methodology address the issue of existing method by reducing dimensionality of the problem,through the classification of individual Web pages into the different categories of the web directory. This issue requires the evaluation of community Web directories according to user preferences. The various types of components of the methodology could be replaced by a number of alternatives. Most importantly, more sophisticated methods for extracting the categories from usage data, in addition to the use of an existing Web directory, would make the mapping of pages to domains and then to categories more accurate and complete. At first module of project we are aiming to use efficient k-means algorithm in order to get efficient results.

The experiment is performed to evaluate the three pattern discovery methods. By this method of personalization we are examining here the percentage of shrinkage of web directory. Average path length is calculated of community web directory and compared the result against original directories. Also we are evaluating here coverage and user gain against another algorithm to remove the problem of local overload.

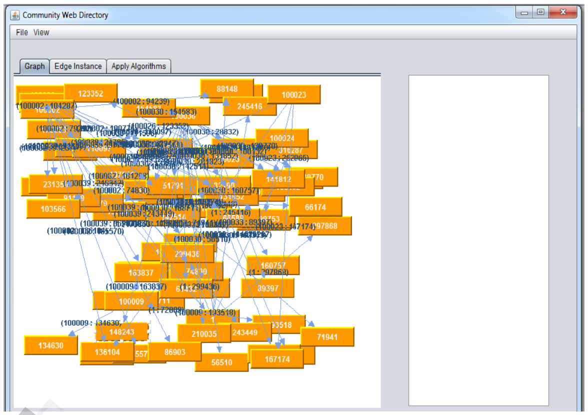
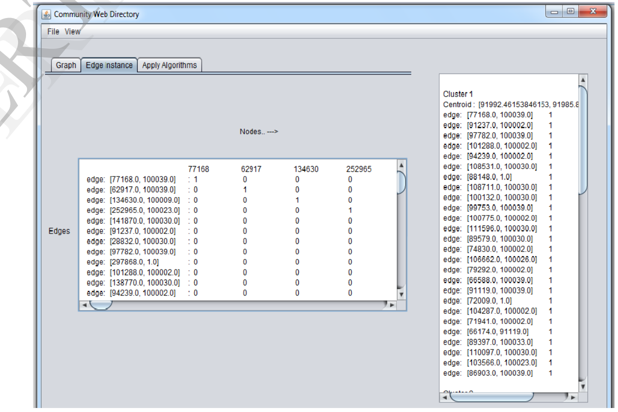


Fig 3: Graph for Dimensionality of user session.



```
                              Nodes.. --->

                                  77168      62917      134630     252965
      edge:  [77168.0, 100039.0]  : 1          0          0          0
      edge:  [62917.0, 100039.0]  : 0          1          0          0
      edge:  [134630.0, 100009.0] : 0          0          1          0
      edge:  [252965.0, 100023.0] : 0          0          0          1
      edge:  [141870.0, 100030.0] : 0          0          0          0
Edges edge:  [91237.0, 100002.0]  : 0          0          0          0
      edge:  [28832.0, 100030.0]  : 0          0          0          0
      edge:  [97782.0, 100039.0]  : 0          0          0          0
      edge:  [297868.0, 1.0]      : 0          0          0          0
      edge:  [101288.0, 100002.0] : 0          0          0          0
      edge:  [138770.0, 100030.0] : 0          0          0          0
      edge:  [94239.0, 100002.0]  : 0          0          0          0
```

```
Cluster 1
Centroid :  [91992.46153846153, 91985.8
edge:  [77168.0, 100039.0]     1
edge:  [91237.0, 100002.0]     1
edge:  [97782.0, 100039.0]     1
edge:  [101288.0, 100002.0]    1
edge:  [94239.0, 100002.0]     1
edge:  [108531.0, 100030.0]    1
edge:  [88148.0, 1.0]          1
edge:  [108711.0, 100030.0]    1
edge:  [100132.0, 100030.0]    1
edge:  [99753.0, 100039.0]     1
edge:  [100775.0, 100002.0]    1
edge:  [111596.0, 100030.0]    1
edge:  [89579.0, 100030.0]     1
edge:  [74830.0, 100002.0]     1
edge:  [106662.0, 100026.0]    1
edge:  [79292.0, 100002.0]     1
edge:  [66588.0, 100039.0]     1
edge:  [91119.0, 100039.0]     1
edge:  [72009.0, 1.0]          1
edge:  [104287.0, 100002.0]    1
edge:  [71941.0, 100002.0]     1
edge:  [66174.0, 91119.0]      1
edge:  [89397.0, 100033.0]     1
edge:  [110097.0, 100030.0]    1
edge:  [103566.0, 100023.0]    1
edge:  [86903.0, 100039.0]     1
```

Fig 4: Clustring of user session.

5. Conclusion

Today world is of internet as the emergence of e-services in the new web era, such as e-commerce, e-learning and e-banking. The internet is used turning web sites into business and increasing competition between them. A web friendly environment is developed by offering personalization of services. The proposed methodology alleviating the problem of information overload by constructing the community web directory to the needs and interest of particular user. With the help of machine learning methods we can construct such directories using clustering and probabilistic approach. As web usages data is diverse and voluminous, it can be reduced by classifying the web pages into classified folders called as directories mapped with user interest. But there is need of future scope to check the robustness of algorithm according to the changing environment and the parameters analysis of the community model.

References

[1] M.Ramkrishna,L.Gowdar,M.Havanur*”WebMining: KeyAccomplishment, Applications And Future Directions”*,International journal of Data Storage and Data EngineerinVol.1, pp 4-9.,2010

[2] D.Pierrakos, G. Paliours, C Papatheodorou, and C. D. Spyropoulos *” Web Usages Mining as a Tool for Personalization : A Survey”*User Modeling and User-Adaption interaction, Vol. 13 , pp.311 372 ,2010

[3] G. Paliouras, C. Papatheodorou, V. Karkaletsis, and C.D.Spyropoulos, *”Discovering User Communities on the Internet Using Unsupervised Machine Learning Techniques”* Interacting with Computers J., vol. 14, no. 6, pp. 761-791, 2002.

[4] D.Pierrakos, Georgios Paliours, *”Personalizing web directories with the Aid of Web Usages Data”* ,IEEE Transactions on Knowledge and Data Engineering,vol.22,no.9,Sep 2010.

[5] D. Pierrakos, G. Paliours, C Papatheodorou, and V. Karkaletis, M Dikaiakos *” Web Community Directories : A New Approach to Web Personalization”* Web Mining : From web to Semantic Web, pp 113-129, Springer,2004.

[6] D. Pierrakos and G. Paliouras, *”Exploiting Probabilistic Latent Information for the Construction of Community Web Directories”*, Proc. 10th IntâĂŹl Conf. User Modeling, L. Ardissono, P. Brna, and A. Mitrovic, eds., pp. 89-98, 2005.

[7] D. Chen, D. Wang, and F. Yu, *”A PLSA-Based Approach for Building User Profile and Implementing Personalized Recommendation”*, Proc. Joint Ninth Asia-Pacific Web Conf. (APWeb ’07) and Eighth Int Conf. Web-Age Information Management (WAIM ’07), pp. 606-613, 2007.

[8] B. Mobasher, R. Cooley, and J. Srivastava,*” Automatic Personalization Based on Web Usage Mining”* Comm. ACM, vol. 43, no. 8, pp. 142-151, 2000.

[9] P. Brusilovsky, A. Kobsa, and W. Neijdl,*” The Adaptive Web, Methods and Strategies of Web Personalization ”*, eds. Springer, 2007.

[10] T. Hofmann,*”Learning What People (Don’t) Want”* Proc. 12th European Conf. in Machine Learning, pp. 214-225, 2001.

[11] G. Xu, Y. Zhang, and Y. Xun,*”Modeling User Behaviour for Web Recommendation Using lda Model”* Proc. IEEE/WIC/ACM Int Conf. Web Intelligence and Intelligent Agent Technology, pp. 529-532, 2008.

[12] W. Chu and S.-T.P. Park,*”Personalized Recommendation on Dynamic Content Using Predictive Bilinear Models”* Proc. 18th Int Conf. World Wide Web (WWW), pp. 691-700, 2009.

[13] X. Jin, Y. Zhou, and B. Mobasher,*”Task-Oriented Web User Modeling for Recommendation”* Proc. 10th International Conf. User Modeling, L. Ardissono, P. Brna, and A. Mitrovic, eds., pp. 109118,2005.

[14] Y. Fu, K. Sandhu, and M. Shih. *”A Generalization-BasedApproach to Clustering ofWeb Usage Sessions”*, InProceedings of the 1999 KDD Workshop on Web Mining,San Diego, CA, vol. 1836 of LNAI,. Springer, 2000, 21-38

[15] M. S. Chen, J. S. Park, and P. S. Yu.*” Efficient Data Mining for Path Traversal Patterns”* Knowledge and Data Engineering, 10(2), 1998, 209-221

[16] A. Joshi and R. Krishnapuram. *”On Mining Web AccessLogs”* In ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery, pages 2000, 63- 69

[17] Bettina Berendt, *”Web usage mining, site semantics, and thesupport of navigation”*, in Proceedings of the Workshop ”WEBKDD 2000 - Web Mining for E-Commerce -Challenges and Opportunities”, 6th ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining, 2000

[18] B. Mobasher, H. Dai, T. Luo, and M. Nakagawa.*” Discovery and Evaluation of Aggregate Usage Profiles for Web Personalization”*. Data Mining and Knowledge Discovery,6(1), 2002, 61-82

[19] P. Berkhin.*”Survey of clustering data mining techniques”*,Springer Berlin Heidelberg, Berlin,2006L. Catledge and J. Pitkow. ”Characterizing browsing behaviors on the World Wide Web”Computer Networks and ISDN Systems ,1995, vol. 27, no. 6, pp. 1065-1073

[20] B. Mobasher, R. Cooley and J. Srivastara. *”Automatic personalization based on Web session clustering”*,Communications of ACM, 2000, vol. 43, no. 8, pp. 142-151

[21] J. Yu . *”General C-means clustering model”* IEEE Transactions on Pattern Analysis and Machine Intelligence, 2005, Vol. 27, No. 8, pp.1197-1211

[22] Y. Fu, K. Sandhu and M. Y. Shih. *”Clustering of Web Users Based on Access Patterns”* Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining Workshop on Web Mining, Springer , 1999, pp. 560-567